Complex Knowledge Base Tools for Natural Language Communication with the Semantic Web

Kamil Ekštein [1], Miloslav Konopík [1], Roman Mouček [1], Tomáš Pavelka [1], Josef Steinberger [1], Roman Tesař [1]

[1] University of West Bohemia, FAV, KIV
Univerzitní 8, 306 14 Pilsen, Czech Republic

E-mail: konopik@kiv.zcu.cz

Keywords: automatic speech recognition, semantic analysis, web page filtration, automatic summarization

Extended Abstract:

In this article we introduce the COT-SEWing (Complex Knowledge Base Tools for Natural Language Communication with the Semantic Web) project. The purpose of the project is to develop a complex base of tools which will remove some of the typical barriers present in communication between human user and computer within the scope of internet access. Our contribution will include methods, algorithms, tools and techniques which will improve the quality of user's experience during web usage, namely:

- voice communication including the making of inquiries in natural language,
- tools for automatic creation of semantic description of utterances in limited domains,
- means for web page and document filtration,
- methods for automatic annotation and summarization of large documents and document collections,
- and others

Voice communication including the making of inquiries in natural language

The LASER (LICS Automatic Speech Extraction/Recognition) recognizer is currently being developed by the Laboratory of Intelligent Information Systems (LICS), University of West Bohemia and will be used to transform the queries and commands spoken in natural language. Since it is not yet possible to recognize general speech (the best recognizers today can deal with dictionaries having about 200 thousand words which is not enough for inflectional languages like Czech) the biggest challenge is how to model the limited domain of internet queries.

To define the domain means to find (and model) a subset of all the possible sentences in the language that are sufficient to cover all commands and inquiries the user can make up. One possibility supported by our recognizer is the usage of context free grammars, but the disadvantage is that the grammars have to be made by hand. Another solution is to use

stochastic language models which can learn the syntax of the language from a set of training examples. Where to find the training data is a question for further research.

Tools for automatic creation of semantic description of utterances in limited domains

The first task within the building tools for automatic creation of semantic description in limited domains is determination of the goal domains and corpora preparation. We have decided to elaborate on domains, in which the answer to a simple question can be easily found in web pages (like weather forecast). Building a corpus is a long-lasting and time-consuming process demanding a large number of participants. There is also a need to map the target group of possible users of the whole system and to find their way of web questioning. We have collected about 27.000 questions in ten domains. The corpus has been built by approximately 450 people (half of them were students, half of them their relatives, friends etc.)

The corpus of typical user utterances has to be annotated. The process of annotation is the process of assigning the meaning to each utterance. The meaning is assigned by a team of human annotators. The meaning of a sentence is represented in a suitable meaning representation. Hence, one step is to propose the suitable representation of the meaning. When the corpus is annotated (at least partially) then a computer algorithm has to be proposed to learn the rules of semantic analysis automatically. The result of this stage of the COT-SEWing project is the algorithm that is capable of automatic semantic analysis. Such an algorithm automatically assigns the meaning to an input utterance.

Means for web page and document filtration

Internet is an immense resource of data. There are billions of documents in various formats – text, image, audio, video, etc. Many of these documents represent useful knowledge that must be extracted out of them first. This extraction (mining) is the subject of a scientific field called Web mining. Recent advances in Web mining have concentrated on content, structure, and usage mining. Both content and structure mining techniques may help us distinguish between relevant and irrelevant Web documents. The former by categorizing into on-topic and off-topic documents, the latter by determining important (authoritative) documents via analysis of their relations to other documents. Of course, the heterogeneous and decentralized nature of the Web causes many useless, harmful or even criminal Web pages to appear. Web mining may also be applied to their detection and elimination.

Methods for automatic annotation and summarization of large documents and document collections

User's experience during web usage can be largely improved by text summarization. Sending a short summary instead of a list of documents as a result of a query would make the web really semantic. We plan to focus on both single- and multi-document summarization in multiple languages. Firstly, we plan to prepare testing corpora. For English we are going to use a standard DUC corpus, for Czech a new summarization corpus will be prepared. Our summarization approach will be based on latent semantic analysis and anaphora resolution [4]. The summaries will be then used by our multilingual searching and extraction system (MUSE) [5] both for better user orientation in retrieved documents and for speeding up the searching process when summaries are indexed for searching.

References

[1] Stochastic Semantic Parsing, PhD thesis exposé, Technical Report No. DCSE/TR-2006-01, University of West Bohemia, 2006.

[2] Tesar R., Poesio M., Strnad V., Jezek K.: "Extending the Single Words-Based Document Model: A Comparison of Bigrams and 2-Itemsets". The 2006 ACM Symposium on Document Engineering (DocEng'06), Amsterdam, Netherlands, ACM press, ISBN 1-59593-515-0, pages 138-146, http://doi.acm.org/10.1145/1166160.1166197, October 2006.

[3] Tesar R., Fiala D., Rousselot F., Jezek K.: "A comparison of two algorithms for discovering repeated word sequences". The 6th International Conference on Data Mining, Text Mining and their Business Applications (Data Mining 2005), Skiathos, Greece, ISBN 1-84564-017-9, pages 121-131, WIT Transaction on Information and Communication Technologies, ISSN 1743-3517, http://library.witpress.com/pages/PaperInfo.asp?PaperID=14997, May 2005.

[4] Josef Steinberger, Mijail A. Kabadjov and Massimo Poesio: Improving LSA-based Summarization with Anaphora Resolution. Proceedings of Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing, The Association for Computational Linguistics, Vancouver, Canada, October 2005, ISBN 1-932432-55-8.

[5] Michal Toman and Josef Steinberger, Karel Ježek: Searching and Summarizing in Multilingual Environment. Proceedings of the 10th International Conference on Electronic Publishing, pp. 257-265, FOI-Commerce, Bansko, Bulgaria, June 2006, ISBN 954-16-0049-9.